A first approach to a computational model for the acquisition of Spanish verbal inflection

Santiago Gualchi (University of Buenos Aires)
santiagogualchi@filo.uba.ar
santiagogualchi@gmail.com

One of the challenges children face during language acquisition consists of completing inflectional paradigms. Chan (2008) and Lignos & Yang (2016) have shown, through corpus studies, that inflectional data available to children are sparse. Only a subset of all possible forms of a word have been found to be present in the linguistic environment, and the distribution of their inflectional values approximates a Zipfian curve. This means that few of them are very common (e.g., third-person singular present indicative, for Spanish verbs), while many are very rare (e.g., second-person plural preterit subjunctive). Similarly, they found that inflectional saturation (i.e., the proportion of possible inflectional forms present in a corpus) is greater for more frequent words, and not even these are saturated to a high degree. In the Spanish corpus examined by Lignos & Yang, for instance, the most saturated word was *decir* 'to say', but saturation was only 72% so over a fourth of all possible inflectional forms of this verb were not present in the data. Therefore, the discovery of regularities between the form of words and their syntactic-semantic properties constitutes a milestone of language acquisition. This way, children can resort to generalisations to generate unattested forms.

Along these lines, previous studies had already stated that children are able to use unlearnt inflected forms. On the one hand, elicitation experiments, such as Berko's (1958), show that children can inflect a word since the earliest exposures even if they had never heard the target form. On the other hand, Marcus et al.'s (1992) large-scale corpus study has found that (aside from omission errors) children acquiring English exhibit a U-shaped development of inflectional morphology which consists of a first phase characterised by a correct use of irregulars followed by a second phase with overregularisations (the use of which decreases gradually during school-age years). This pattern cannot be explained as a result of exposure to said forms as they are unlikely to be present in the child's linguistic environment. Still, analogical errors (or overirregularisations) are almost nonexistent (Xu & Pinker, 1995).

Different mechanisms have been proposed to account for this behaviour. On one end, empiricist approaches seek to provide an explanation on the basis of general-purpose learning mechanisms (e.g., Bybee, 1995; Rumelhart and McClelland, 1986; Tomasello, 2003). These models are based on information processing and the discovery of quantitative associations and patterns between words and their properties. On the other end, rationalist proposals argue that general-domain processes fail to provide a suitable explanation of children's linguistic performance, and claim that there are language-specific learning mechanisms and/or innate linguistic knowledge mediating (e.g., Pinker, 1991; Marcus et al., 1992, chap. 8). These models rely on dual-route processing: regular forms are computed through symbolic rules applied to the base (e.g., "add *-ed*"), while irregular forms are stored by rote in the lexicon and, therefore, can be directly retrieved from memory. More recently, it has been suggested that a proper answer to this problem is to be found somewhere in the middle (e.g., Allen & Behrens, 2019). Along these lines, models such as Yang's (2002) rely heavily on associative learning for the pairing of words and inflectional rules while still keeping a default rule that applies every time the system fails to retrieve an irregular one.

Most models of inflectional morphology acquisition have sought to provide an answer to the problem of English past tense morphology. Nevertheless, this debate does not tackle the vast complexity of inflectional morphology crosslinguistically. As Tomasello (2003) points out, English

past tense has only one productive morpheme (*-ed*) which is a suffix relatively easy to concatenate with verb stems while irregulars typically involve phonologically-conditioned stem changes. For this reason, other systems, such as the German nominal plural or the Polish morphological case have been explored. As for Spanish verbal inflection, this system is of interest for several reasons: (1) Spanish verbal inflection is usually analysed as comprising two morphemes: one for tense, aspect and mood, and another for subject agreement (e.g., RAE & Asale, 2009, sec. 4.1a); (2) there are three conjugational classes (I, *-a-*; II, *-e-*; and III, *-i-*); (3) irregulars are formed through a large number of morphological processes from semi-regular phonological changes (such as diphthongisation) to suppletion; (4) there are tens of morphological forms for each verb (the exact number depending on dialectal variation); and (5) the availability of forms in the linguistic environment is asymmetrical (e.g., indicative is more common than subjunctive).

Children's behaviour during the acquisition of Spanish morphology has some resemblance to that of children acquiring English. Pérez Pereira & Singer (1984), for example, found through elicitation experiments that children would overregularise irregulars and commit omission errors. They also noticed that children would overgeneralise some inflections of conjugational class I to conjugational classes II and III. These and other characteristics, such as the observation of conjugational classes for determining the inflection of a verb (Brovatto & Ullman, 2005), have been interpreted as evidence of a qualitative distinction between regular and irregular verbs in Spanish. Nevertheless, Eddington (2009) confirmed that those properties can be reproduced with single-route models. In light of this, I present the architecture of a connectionist network intended to extract patterns from observed verb forms and generalise to unattested cases. To do this, the model is fed with the phonological and the syntactic content of verbs but no semantic-specific properties are included (in compliance with the T-model of the Language Faculty). This network is inspired in the design of autoencoders (see, for example, Wang et al., 2016), which I suggest can help model both the identification of stems and the generation of inflected forms. Finally, I discuss some strengths, challenges and limitations of such a model.

References

Allen, S. E. M. & H. Behrens (2019). Insights into understanding human language from children's acquisition of Morphology and Syntax: A historical and current perspective on central questions in the field. In P. Hagoort (Ed.), *Human language: From genes and brains to behavior* (pp. 127–145). Cambridge, MA: The MIT Press. || **Berko, J. (1958)**. The child's learning of English morphology. *WORD*, 14(2–3), 150–177. DOI: 10.1080/00437956.1958.11659661. || **Brovatto, C. & M. T. Ullman (2005)**. The mental representation and processing of Spanish verbal morphology. In D. Eddington (Ed.), *Selected Proceedings of the 7th Hispanic Linguistics Symposium* (pp. 98-105). Somerville, MA: Cascadilla Proceedings Project. || **Bybee, J. L. (1995)**. Regular morphology and the lexicon. *Language and cognitive processes*, 10, 425–445. DOI: 10.1080/01690969508407111. || **Chan, E. (2008)**. *Structures and distributions in morphology learning* (PhD thesis). Pennsylvania: University of Pennsylvania. || **Eddington, D. (2009)**. Spanish verbal inflection: A single- or dual-route system? *Linguistics*, 47(1), 173-199. DOI: 10.1515/LING.2009.006. || **Lignos, C. & C. Yang (2016)**. Morphology and language acquisition. In A. Hippisley & G. Stump (Eds.), *The Cambridge handbook of morphology* (pp. 765–791). Cambridge: Cambridge University Press. DOI: 10.1017/9781139814720.027. || **Marcus, G. F. et al. (1992)**. Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, 57(4). DOI: 10.2307/1166115. || **Pérez Pereira, M. & D. Singer (1984)**. Adquisición de morfemas del español. *Infancia y aprendizaje*, 27/28, 205-221. || **Pinker, S. (1991)**. Rules of language. *Science*, 253, 530–535. DOI: 10.1126/science.1857983. || **RAE & Asale (2009)**. La flexión verbal. In *Nueva gramática de la lengua española*. Vol. Morfología, Sintaxis I (pp. 181-335). Madrid: Espasa Libros. || **Rumelhart, D. E. & J. L. McClelland (1986)**. On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart & The PDP Research Group, *Parallel distributed processing* Vol. 2 (pp. 216–271). Cambridge, MA: The MIT Press. || **Tomasello, M. (2003)**. *Constructing a language*. Cambridge, MA: Harvard University Press. || **Wang, Y. et al. (2016)**. Auto-encoder based dimensionality reduction. *Neurocomputing*, 184, 232-242. DOI: 10.1016/j.neucom.2015.08.104. || **Xu, F. & S. Pinker (1995)**. Weird past tense forms. *Journal of Child Language*, 22, 531–556. DOI: 10.1017/S0305000900009946. || **Yang, C. (2002)**. *Knowledge and learning in natural language*. Boston: Springer. DOI: 10.1007/978-1-4419-1428-6_837.